Design 1: k-nearest neighbors with default parameters.
k-NN separates data into several classes, in order to predict the classification of a variable, so we believed it would be one of the most effective algorithms for the training dataset. It provides a starting point and basic result that we could build from.

Design 2: Random Forest model
It is known for its accuracy with large data as well as data containing missing values. The model used 100 trees with a maximum depth of 10, without applying pruning or pre-pruning. With such a large set of data, we felt this would end up being an important part of model, due to the Random Forest model increasing in accuracy as the model grows, as long as it doesn't begin to overfit. It performed slightly better than k-NN with default parameters and struggled with DNA/RNA sensitivity as well as nonDRNA specificity.

Design 3: combined three models, k-NN with default parameters, k-NN with k = 50, and Random Forest, by using the Vote operator in RapidMiner.
I chose to combine two values for k due to its increase in overall accuracy when increasing k, and Random Forest due to Its high overall accuracy from the start. The combination of these three models did result in a higher overall accuracy, but for the case of nonDRNA, a decrease in sensitivity, and an increase in sensitivity for DNA, RNA, and DRNA. It also created a large increase in specificity for nonDRNA, but a slight decrease in specificity for DNA, RNA, and DRNA.

Table 1 Key:
Design 1: k-NN (default)
Design 2 : Random Forest
Design 3 : k-NN(default) & k-NN(k=50) & Random Forest
Best Design :

Results (see Evaluation of Predictions section):

Table 1. Summary of results based on the 5-fold cross validation on the training dataset.

Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	8.9	10.0	2.6	28.9	
	Specificity	99.9	95.6	99.5	95.4	
	PredictiveACC	95.6	91.4	94.9	95.0	
	MCC	**0.255**	**0.056**	**0.056**	**0.090**	
RNA	*Sensitivity*	44.1	21.2	12.0	52.8	
	Specificity	99.0	95.4	98.9	94.3	
	PredictiveACC	96.0	90.6	93.5	93.8	
	MCC	**0.549**	**0.127**	**0.144**	**0.147**	
DRNA	*Sensitivity*	0.0	0	0	0	
	Specificity	100.0	100.0	100.0	99.7	
	PredictiveACC	99.7	100.0	99.7	99.7	
	MCC	**0**	**0**	**0**	**0**	
nonDRNA	*Sensitivity*	99.1	92.0	98.7	90.2	
	Specificity	28.7	17.3	8.0	52.9	
	PredictiveACC	91.6	84.6	89.2	89.7	
	MCC	**0.443**	**0.097**	**0.143**	**0.163**	
averageMCC		**0.312**	**0.070**	**0.086**	**0.100**	
accuracy		91.4	83.9	89.0	89.4	

Confusion Matrix: